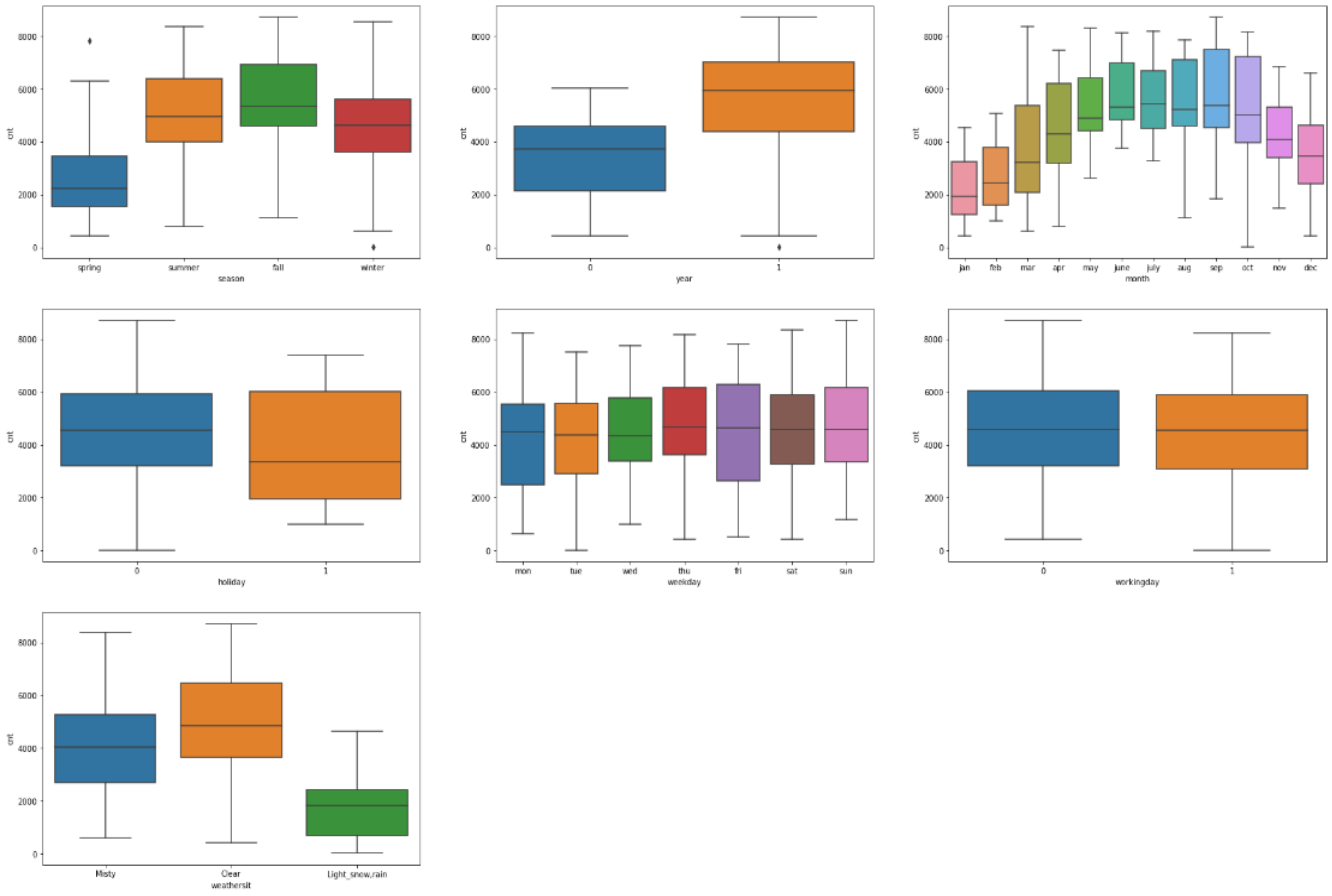
Linear Regression Subjective Questions

- **Srilathaa Vasu**

Assignment based subjective questions:

1. **From your analysis of the categorical variables from the dataset, what could you infer about their effect on the dependent variable? (3 marks)**

Following are the categorical variables observed in the dataset - season, year, month, holiday, weekday, working day and weather sit. I have analyzed these variables using the boxplot.



- Season - Most of the bike booking happened in fall and summer season with a median of over 5000 bookings. Season can be one of the good predictor for the dependent variable.
- Year - There were more demand in 2019 compared to 2018.
- Month - September month has the highest demand with around 7000 maximum bookings. The number of ride count increases between May and October. These months are summer and fall season in US.

- Holiday - Most of the bookings happened when it is not a holiday. Demand has decreased when there is a holiday.
- Weekdays - Weekday variable can have some or no influence towards predictor since they show similar trend with a median of over 5000 bookings.
- Working Day - Majority of the bike booking were happening in ‘workingday’ with a median of close to 5000 booking.
- Weather sit - The clear weather has the highest demand followed by misty weather.

2. Why is it important to use drop_first=True during dummy variable creation? (2 marks)

It is advisable to use drop_first=True, as it helps in reducing the extra column created during dummy variable creation. Hence it reduces the correlations created among dummy variables.

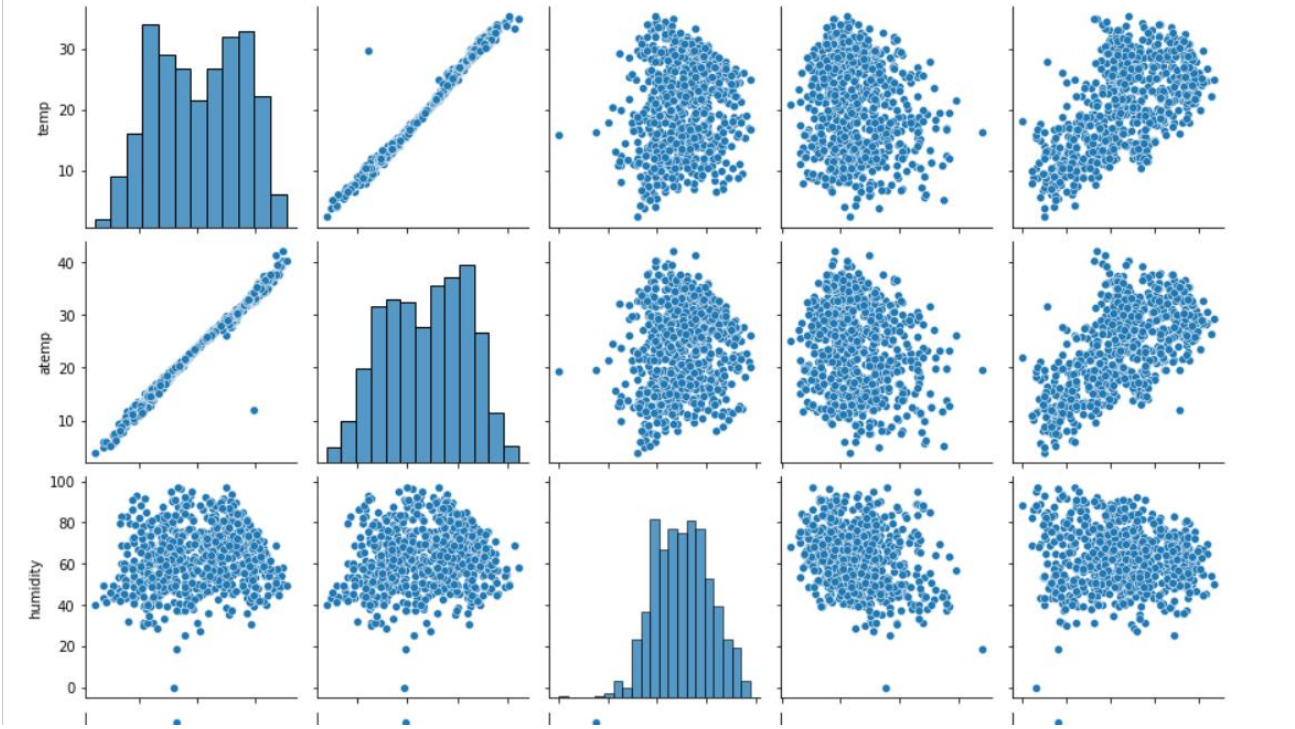
Syntax : drop_first : bool, default False
Whether to get k-1 dummies out of k categorical levels by removing the first level.

If we don’t drop the first column then the dummy variables will be correlated. This may affect some models adversely and the effect is stronger when the cardinality is smaller. Let’s say we have 3 types of values in a categorical column and we want to create a dummy variable for that column. If one variable is not A and B, then it means that it is C. So we don’t need the third variable to identify C.

3. Looking at the pair-plot among the numerical variables, which one has the highest correlation with the target variable? (1 mark)

From the pair plot, we can observe that temp and atemp have the highest correlation with the target variable cnt.

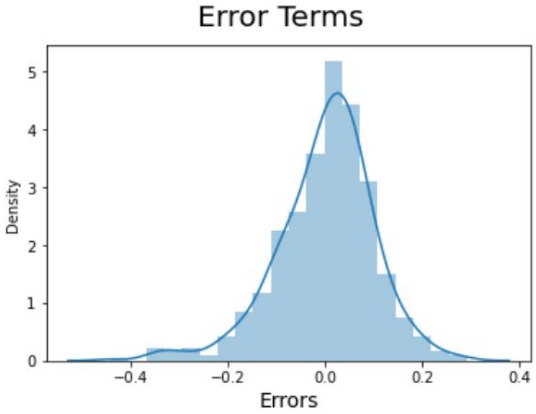
```
Text(0.5, 1.0, 'Plot for continuous numeric variables')

<Figure size 720x432 with 0 Axes>
```



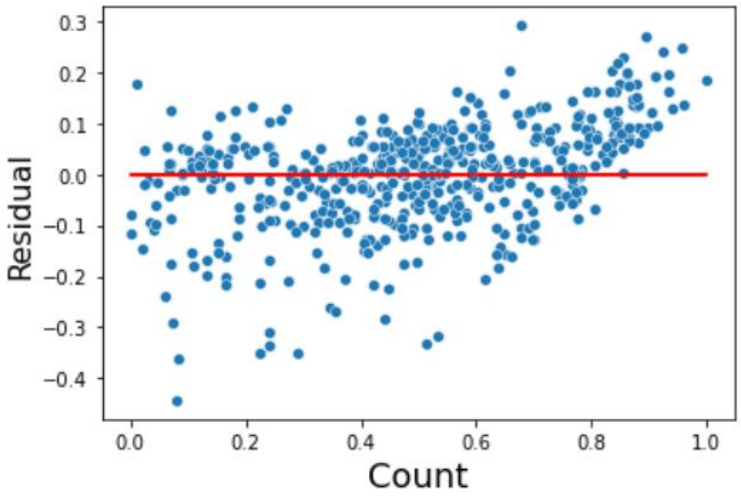
4. How did you validate the assumptions of Linear Regression after building the model on the training set? (3 marks)

Based on the following assumptions, I have validated the assumptions of my Linear Regression model.

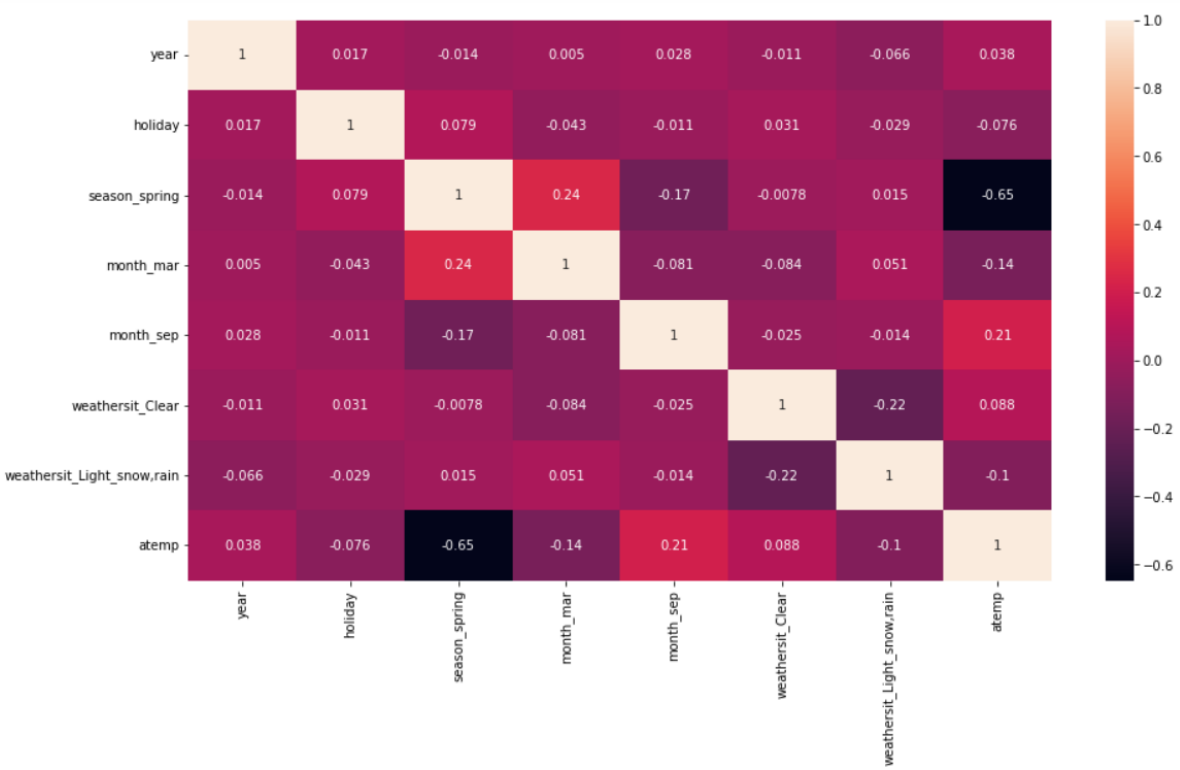
- Normality of error terms: Error terms should be normally distributed with mean zero.



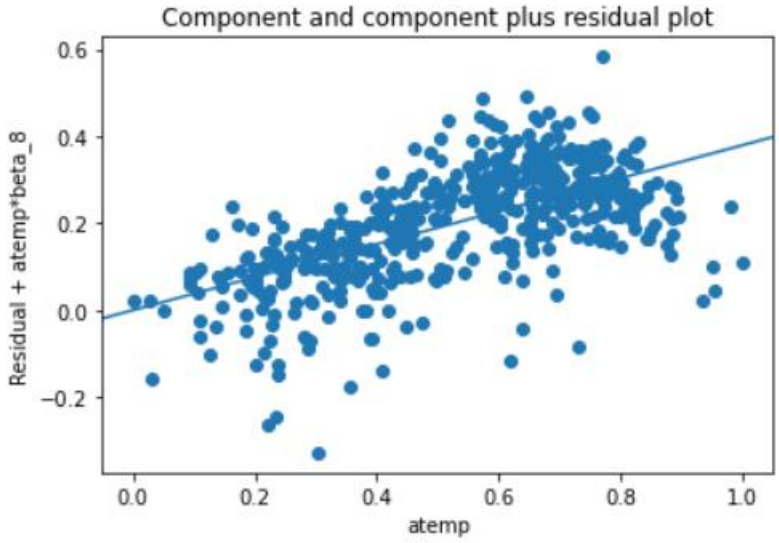
- Homoscedasticity: There should be no visible patterns in the residual values.



- Multi collinearity: There should be insignificant multi collinearity among variables.



- Linear Relationship: The relationship between the model and the predictor variables needs to be linear.



5. **Based on the final model, which are the top 3 features contributing significantly towards explaining the demand of the shared bikes? (2 marks)**

The top 3 features which contribute towards explaining the demand of the shared bikes are as follows:

- Atemp
- Year
- Season

General Subjective questions:

1. **Explain the linear regression algorithm in detail. (4 marks)**

A simple linear regression model attempts to explain the relationship between a dependent variable and an independent one using a straight line. The independent variable is known as the predictor variable. The dependent variable are also known as the output variables.

Linear relationship between variables means that when the value of one or more independent variables will change (increase or decrease), the value of dependent variable will also change accordingly (increase or decrease). Mathematically the relationship can be represented with the help of following equation

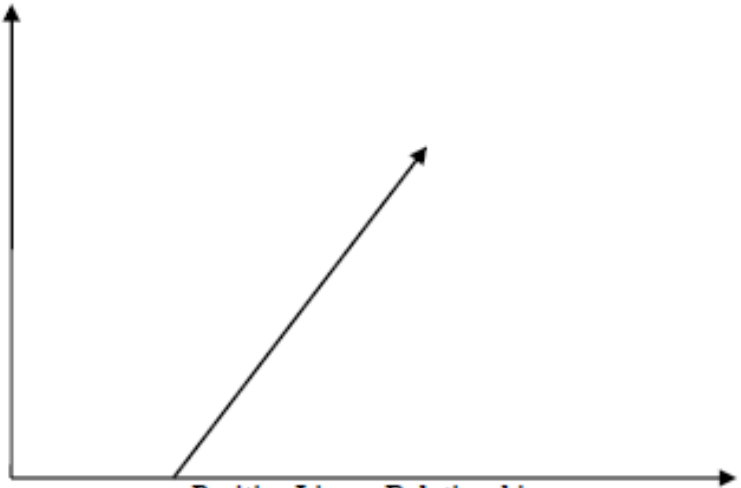
$$Y = mX + c$$

Here, Y is the dependent variable we are trying to predict. X is the independent variable we are using to make predictions. m is the slope of the regression line which

represents the effect X has on Y c is a constant, known as the Y-intercept. If X = 0, Y would be equal to c.

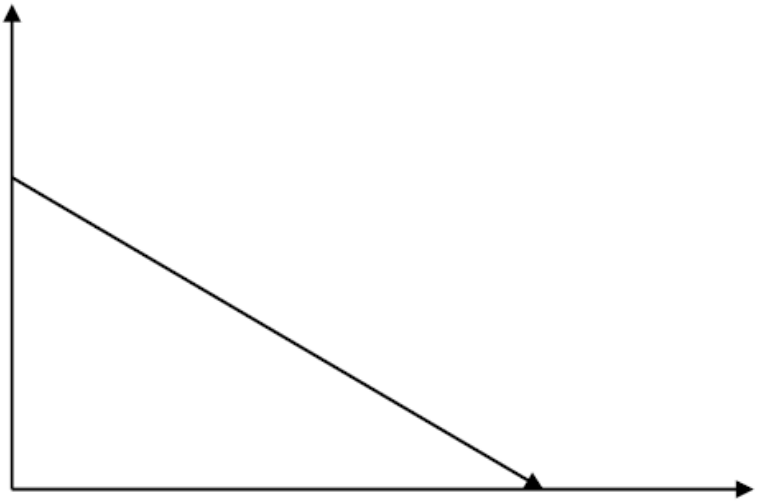
Furthermore, the linear relationship can be positive or negative in nature as explained below

a. Positive Linear Relationship:
   - A linear relationship will be called positive if both independent and dependent variable increases. It can be understood with the help of following graph,



Positive Linear Relationship

b. Negative Linear relationship:
   - A linear relationship will be called positive if independent increases and dependent variable decreases. It can be understood with the help of following graph,



Negative Linear Relationship

Linear regression is of the following two types

I. Simple Linear Regression
II. Multiple Linear Regression

Assumptions:

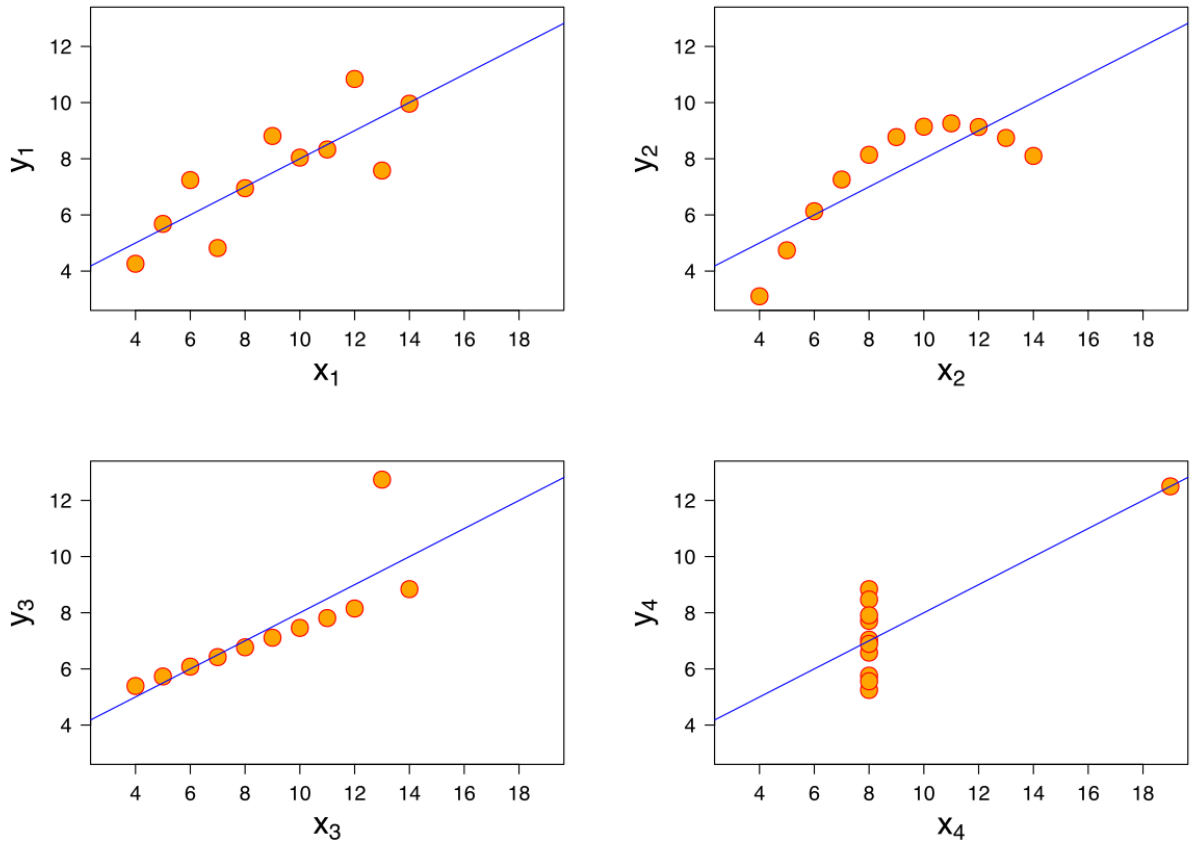
The following are some assumptions about dataset that is made by Linear Regression model

- Multi-collinearity:
  Linear regression model assumes that there is very little or no multi-collinearity in the data. Basically, multi- collinearity occurs when the independent variables or features have dependency in them.
- Auto-correlation:
  Another assumption Linear regression model assumes is that there is very little or no auto-correlation in the data. Basically, auto-correlation occurs when there is dependency between residual errors.
- Relationship between variables:
  Linear regression model assumes that the relationship between response and feature variables must be linear.
- Normality of error terms:
  Error terms should be normally distributed.
- Homoscedasticity:
  There should be no visible pattern in residual values.

2. Explain the Anscombe's quartet in detail. (3 marks)

Anscombe's Quartet was developed by statistician Francis Anscombe. It comprises four datasets, each containing eleven (x, y) pairs. The essential thing to note about these datasets is that they share the same descriptive statistics. But things change completely, and I must emphasize COMPLETELY, when they are graphed. Each graph tells a different story irrespective of their similar summary statistics.

The summary statistics show that the means and the variances were identical for x and y across the groups: • Mean of x is 9 and mean of y is 7.50 for each dataset. • Similarly, the variance of x is 11 and variance of y is 4.13 for each dataset • The correlation coefficient (how strong a relationship is between two variables) between x and y is 0.816 for each dataset When we plot these four datasets on an x/y coordinate plane, we can observe that they show the same regression lines as well but each dataset is telling a different story:



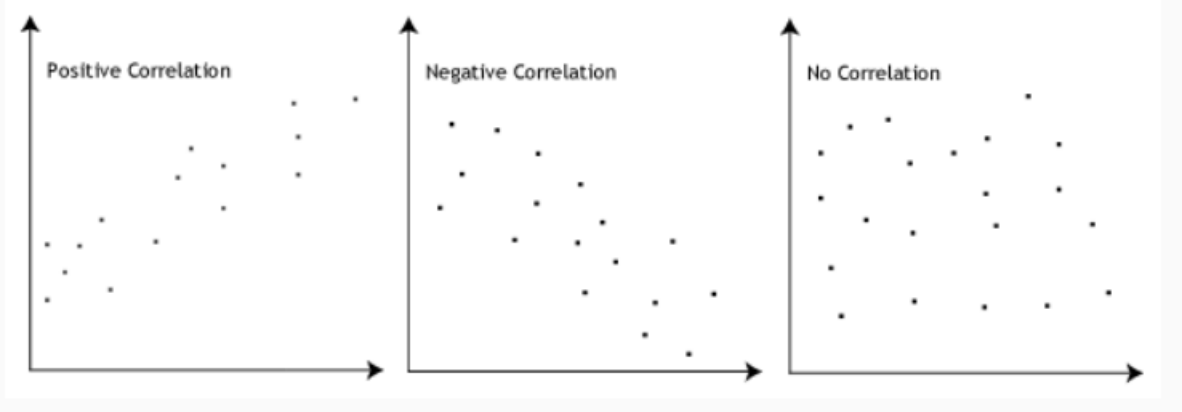
- Dataset I appears to have clean and well-fitting linear models.
- Dataset II is not distributed normally.
- In Dataset III the distribution is linear, but the calculated regression is thrown off by an outlier.
- Dataset IV shows that one outlier is enough to produce a high correlation coefficient. This quartet emphasizes the importance of visualization in Data Analysis. Looking at the data reveals a lot of the structure and a clear picture of the dataset.

3. What is Pearson's R? (3 marks)

Pearson's r is a numerical summary of the strength of the linear association between the variables. If the variables tend to go up and down together, the correlation coefficient will be positive. If the variables tend to go up and down in opposition with low values of one variable associated with high values of the other, the correlation coefficient will be negative.

The Pearson correlation coefficient, r, can take a range of values from +1 to -1. A value of 0 indicates that there is no association between the two variables. A value greater than 0 indicates a positive association; that is, as the value of one variable increases, so does the value of the other variable. A value less than 0 indicates a negative association; that is, as

the value of one variable increases, the value of the other variable decreases. This is shown in the diagram below:



4. What is scaling? Why is scaling performed? What is the difference between normalized scaling and standardized scaling? (3 marks)

Feature Scaling is a technique to standardize the independent features present in the data in a fixed range. It is performed during the data pre-processing to handle highly varying magnitudes or values or units. If feature scaling is not done, then a machine learning algorithm tends to weigh greater values, higher and consider smaller values as the lower values, regardless of the unit of the values.

Example: If an algorithm is not using feature scaling method then it can consider the value 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithm will give wrong predictions. So, we use Feature Scaling to bring all values to same magnitudes and thus, tackle this issue.

S.No.	Normalization	Standardization
1.	Minimum and maximum value of features are used for scaling	Mean and standard deviation is used for scaling.
2.	It is used when features are of different scales.	It is used when we want to ensure zero mean and unit standard deviation.
3.	Scales values between [0, 1] or [-1, 1].	It is not bounded to a certain range.
4.	It is really affected by outliers.	It is much less affected by outliers.
5.	Scikit-Learn provides a transformer called MinMaxScaler for Normalization.	Scikit-Learn provides a transformer called StandardScaler for standardization.
6.	This transformation squishes the n-dimensional data into an n-dimensional unit hypercube.	It translates the data to the mean vector of original data to the origin and squishes or expands.
7.	It is useful when we don't know about the distribution	It is useful when the feature distribution is Normal or Gaussian.

8.	It is a often called as Scaling Normalization	It is a often called as Z-Score Normalization.

Formula of Normalized scaling:

$$x = \frac{x - min(x)}{max(x) - min(x)}$$

Formula of Standardized scaling:

$$x = \frac{x - mean(x)}{sd(x)}$$

5. **You might have observed that sometimes the value of VIF is infinite. Why does this happen? (3 marks)**

VIF – Variance Inflation Factor

The value of VIF is calculated by the below formula:

$$VIF = \frac{1}{1 - R^2}$$

Where, 'i' refers to the ith variable.

If there is perfect correlation, then VIF = infinity. A large value of VIF indicates that there is a correlation between the variables. If the VIF is 4, this means that the variance of the model coefficient is inflated by a factor of 4 due to the presence of multicollinearity.

When the value of VIF is infinite it shows a perfect correlation between two independent variables. In the case of perfect correlation, we get R-squared (R2) =1, which lead to 1/ (1-R2) infinity. To solve this we need to drop one of the variables from the dataset which is causing this perfect multicollinearity.

6. **What is a Q-Q plot? Explain the use and importance of a Q-Q plot in linear regression. (3 marks)**

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution.

Use of Q-Q plot:

A q-q plot is a plot of the quantiles of the first data set against the quantiles of the second dataset. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.3 (or 30%) quantile is the point at which 30% percent of the data fall below and 70% fall above that value. A 45-degree reference line is also plotted. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions.

Importance of Q-Q plot:

When there are two data samples, it is often desirable to know if the assumption of a common distribution is justified. If so, then location and scale estimators can pool both data sets to obtain estimates of the common location and scale. If two samples do differ, it is also useful to gain some understanding of the differences. The q-q plot can provide more insight into the nature of the difference than analytical methods such as the chi-square and Kolmogorov-Smirnov 2-sample tests.